# Latent Factor Regression Models for Grouped Outcomes

D. B. WOODARD, T. M. T. LOVE, S. W. THURSTON,

D. RUPPERT, S. SATHYANARAYANA AND S. H. SWAN

We consider models for the effect of exposure on multiple outcomes, where the outcomes are nested in domains. We show that random effect models for this nested situation fit into a standard factor model framework, which leads us to view the modeling options as a spectrum between parsimonious random effect multiple outcomes models and more general continuous latent factor models. We introduce a set of models along this spectrum that extend an existing random effect model for multiple outcomes nested in domains. We characterize the tradeoffs between parsimony and flexibility in this set of models, applying them to both simulated data and data relating phthalate exposure to infant anthropometry.

**Key words:** factor analysis, Bayesian, multiple outcomes, phthalates.

---

Corresponding Author: Dawn B. Woodard, School of Operations Research and Information Engineering, 206 Rhodes Hall, Cornell University, Ithaca, NY, 14850. David Ruppert: School of Operations Research and Information Engineering and Department of Statistical Science at Cornell. Tanzy Love and Sally W. Thurston: Department of Biostatistics and Computational Biology at the University of Rochester. Sheela Sathyanarayana: Departments of Occupational and Environmental Health Science and Pediatrics at the University of Washington. Shanna H. Swan: Department of Preventive Medicine at Mount Sinai School of Medicine. Research partly supported by NSF grant CMMI-0926814.

# 1 INTRODUCTION

Multiple-outcome regression models pool information across related outcome variables; this leads to higher power to detect a significant exposure effect than fitting separate regression models (Thurston, Ruppert and Davidson 2009). Such joint models are popular, for instance, in epidemiological studies, which often have multiple measures of physiological or psychological health and attempt to detect small but important effects of low-dose exposure on those outcomes. In such contexts it is crucial to use the available information as efficiently as possible.

There are two general approaches to modeling the effect of an exposure variable on multiple correlated outcomes. One approach models the exposure effect on the outcomes directly (Sammel, Lin, and Ryan 1999; Lin et al. 2000; Coull, Hobert, Ryan, and Holmes 2001; Roy, Lin, and Ryan 2003; Thurston et al. 2009) and induces correlations between outcomes with random effects. Another approach, called the continuous latent factor approach, introduces one or more continuous latent variables that are manifested by the multiple outcomes (Dunson 2000; Budtz-Jorgensen, Keiding, Grandjean and Weihe 2002; Muthén 2002; Budtz-Jorgensen, Keiding, Grandjean, Weihe and White 2003; Sanchez, Budtz-Jorgensen, Ryan and Hu 2005). The direct modeling approach includes the case where the outcomes are nested in domains (Thurston et al. 2009). This is a common situation in epidemiology studies, where one is interested in the effect of exposure on a set of outcomes within domains such as motor function, intelligence, and attention.

We show that the random effect model of Thurston et al. (2009) for multiple outcomes nested in domains is a special case of the continuous latent factor model framework given in Sanchez et al. (2005). This is not surprising since the latter is extremely general, and non-identifiable in the unrestricted case. However, expressing the model of Thurston et al. (2009) in this way suggests extensions and allows us to view the options for modeling grouped outcomes as a spectrum between parsimonious, but less flexible, random effect models and highly parameterized, but more flexible, latent factor models. We introduce a set of models along this spectrum and show that they are identifiable (Section 2). We characterize the tradeoffs between parsimony and flexibility in

this set of models by applying them to both simulated data (Section 3) and data relating phthalate exposure to infant anthropometry (Section 4).

Phthalates are synthetic chemicals that have been found to have toxic effects on developing endocrine, immune, and reproductive systems in animal studies. In human male infants, prenatal exposure has been linked to reduced anogenital distance, a sexually dimorphic trait, and postnatal breast milk exposure has been associated with altered reproductive hormone concentration (Swan et al. 2005; Sathyanarayana, Calafat, Liu, and Swan 2008). There is a great deal of interest in whether phthalate exposure is associated with changes in other sexually dimorphic features in infants, such as head circumference, weight, and measures of body fatness like skinfold thickness. We analyze data from the Study for Future Families (Swan et al. 2003) to address this question. While separate regression models fit to the different outcomes have not detected a significant link to phthalate exposure, multiple-outcomes models like the ones we address here pool information across related outcomes and thus have a better ability to detect such relationships; we investigate the possibility of a phthalate effect using these models.

Both the simulation study and the phthalates analysis give evidence in support of using the most general model out of those that we introduce. In the simulation study it has excellent accuracy in point estimation of the outcome-specific exposure and covariate effects and acceptable accuracy in estimation of the remaining parameters, regardless of which model we use to simulate the data. We do find additional improvement in accuracy associated with using one of our more parsimonious models when the data are drawn from that model. We also find a loss of accuracy for the outcome-specific effects associated with using more parsimonious models when the data are drawn from a more general model. Even so, all models we examined in the simulation study estimated these effects with low error, even when the data were simulated under another model.

For the phthalates data, we do not find a significant effect of phthalate exposure on any of the outcomes, so we use the covariate effects to illustrate the differences between the models. The point estimates show differing degrees of shrinkage for the different models. Interestingly, for this

dataset at least as many significant covariate effects are found in our most general model as in all but the simplest model. This is despite the fact that the interval estimates are wider in the most general model. So in this dataset we do not see low power to detect covariate effects in the most general model, again supporting its use.

# 2 MODELING GROUPED OUTCOMES

First we describe the random effect model of Thurston et al. (2009) for multiple outcomes nested in domains, and show that it is a type of continuous latent factor model in the sense of Sanchez et al. (2005), with one factor for each domain. The continuous latent factor model is itself a special case of a structural equation model (SEM: cf. Sanchez et al. 2005). So we traverse a spectrum from parsimony to flexibility as we go from random effect models to latent variable models to SEMs, and the model-choice decision is not between a SEM and a random effect model, but rather about the appropriate amount of parsimony when considering model restrictions within the SEM framework.

Denote the outcome measurements by $Y_{ij}$ for subjects $i = 1,\ldots,n$ and outcomes $j = 1,\ldots,p$. Although we focus on the case of continuous outcomes, one can handle the discrete case by use of the generalized linear model framework. The outcomes are grouped into domains $d(j) \in \{1,\ldots,d\}$, which are defined to contain strongly positively correlated outcomes. Denote the covariates by the length-$r$ vector $\mathbf{Z}_i$, and the (observed) exposure by $\eta_i$. In accordance with the epidemiology literature we distinguish notationally between these two sets of predictors, although they will be modeled in identical fashion; one can also drop $\eta_i$ in the following models in order to obtain a single undifferentiated set of predictors that includes exposure.

The model of Thurston et al. (2009) extends the linear mixed model approach to borrow information across outcomes and domains while estimating the exposure effect. It provides shrinkage of this effect across domains and across outcomes within a domain, and has higher power to detect

an effect than separate regression models. Their original model is as follows, where the outcome variables, exposure, and all covariates are assumed to be standardized, and where the notation $\overset{\text{ind}}{\sim}$ indicates that the random effects are independently distributed.

$$Y_{ij} = (b_\eta + b_{\text{D},\eta,d(j)} + b_{o,\eta,j})\eta_i + (\mathbf{b}_z + \mathbf{b}_{\text{D},z,d(j)} + \mathbf{b}_{o,z,j})\mathbf{Z}_i + q_i + q_{i,d(j)} + e_{ij} \tag{2.1}$$

where $b_\eta$ is a common exposure effect, $b_{\text{D},\eta,k} \overset{\text{ind}}{\sim} N(0,\tau^2_{\text{D}})$ for $k = 1,\ldots,d$ is a domain-specific exposure effect, $b_{o,\eta,j} \overset{\text{ind}}{\sim} N(0,\tau^2_o)$ is an outcome-specific exposure effect, $\mathbf{b}_z$ is a vector of overall covariate effects, $b_{\text{D},z,k,\ell} \overset{\text{ind}}{\sim} N(0,\tau^2_{\text{D},\ell})$ is a domain-specific covariate effect for the $\ell$th covariate, $b_{o,z,j,\ell} \overset{\text{ind}}{\sim} N(0,\tau^2_{o,\ell})$ is an outcome-specific covariate effect for the $\ell$th covariate, $q_i \overset{\text{ind}}{\sim} N(0,\tau^2_q)$ is a subject-specific random effect, $q_{i,k} \overset{\text{ind}}{\sim} N(0,\tau^2_{q,k})$ is a subject-domain effect, and $e_{ij} \overset{\text{ind}}{\sim} N(0,\sigma^2_j)$ is the residual error. The subject random effect $q_i$ captures the situation where all outcome measures are positively correlated even after accounting for covariates and exposure. The subject-domain effect $q_{i,k}$ captures additional correlation between outcomes within a domain. No intercept parameters are included by Thurston et al. (2009) in model (2.1) due to centering of outcomes, exposure, and all covariates. The class of models proposed by Thurston et al. (2009) allows the domain-specific exposure and covariate effects, $b_{\text{D},\eta,d(j)}$ and $\mathbf{b}_{\text{D},z,d(j)}$, to be treated either as random effects as in (2.1) or as fixed effects.

In contrast with (2.1), a continuous latent factor model induces correlation between related outcomes by assuming that they are all manifestations of a set of common unmeasurable (latent) variables. The general form of a continuous latent factor (CLF) model is the following (Sammel and Ryan 1996; Muthén 2002; Sanchez et al. 2005), where we take the number of factors equal to the number of domains:

$$\begin{aligned}\mathbf{Y}_i &= \boldsymbol{\alpha} + \boldsymbol{\beta}_{o,\eta}\eta_i + \boldsymbol{\beta}_{o,z}\mathbf{Z}_i + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\xi}_i &= \boldsymbol{\beta}_{\text{D},\eta}\eta_i + \boldsymbol{\beta}_{\text{D},z}\mathbf{Z}_i + \mathbf{B}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i.\end{aligned} \tag{2.2}$$

Here, $\mathbf{Y}_i$ is the length-$p$ vector of outcomes for the $i$th subject, $\boldsymbol{\alpha}$ is a length-$p$ vector of intercepts, $\boldsymbol{\beta}_{o,\eta}$ and $\boldsymbol{\beta}_{o,z}$ are $p \times 1$ and $p \times r$ matrices of regression coefficients, $\boldsymbol{\Lambda}$ is a $p \times d$ matrix of factor

loadings, $\boldsymbol{\xi}_i$ is a length-$d$ vector of latent factors, $\boldsymbol{\varepsilon}_i$ is a length-$p$ vector of independent residuals such that $\varepsilon_{ij} \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$, $\boldsymbol{\beta}_{\text{D},\eta}$ and $\boldsymbol{\beta}_{\text{D},z}$ are $d \times 1$ and $d \times r$ matrices of regression coefficients, $\boldsymbol{\zeta}_i$ is a length-$d$ vector such that $\zeta_{i,k} \overset{\text{ind}}{\sim} N(0, \tau^2_{\zeta,k})$, and $\mathbf{B}$ is a $d \times d$ matrix with zero diagonal elements and $(\mathbf{I} - \mathbf{B})$ invertible. Without further restrictions this model is non-identifiable.

To see that (2.1) is a special case of (2.2), specify the factor loadings matrix $\boldsymbol{\Lambda}$ so that the latent traits $\boldsymbol{\xi}_i$ correspond to the outcome domains. To do this, the nonzero elements of $\boldsymbol{\Lambda}$ should be the elements $(j, d(j))$ for each $j$, which we call $\lambda_j$. For example, if there are two domains and four outcomes with $d(1) = d(2) = 1$ and $d(3) = d(4) = 2$ then

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & \lambda_4 \end{pmatrix}^T.$$

For identifiability, it is common practice in factor analysis to set $\lambda_j = 1$ for the first outcome measurement $j$ in each domain (Sanchez et al. 2005).

Next, a parsimonious way to induce $\mathbf{B}$ is to take the second line of (2.2) to be $\boldsymbol{\xi}_i = \boldsymbol{\beta}_{\text{D},\eta}\eta_i + \boldsymbol{\beta}_{\text{D},z}\mathbf{Z}_i + \phi_i + \boldsymbol{\psi}_i$, where $\phi_i \overset{\text{ind}}{\sim} N(0, \tau^2_\phi)$ is a scalar and $\boldsymbol{\psi}_i$ is a length-$d$ vector such that $\psi_{ik} \overset{\text{ind}}{\sim} N(0, \tau^2_{\psi,k})$. This is shown in Appendix A to be a special case of (2.2). Using these specifications,

$$\begin{aligned} Y_{ij} &= \alpha_j + \beta_{o,\eta,j}\eta_i + \boldsymbol{\beta}_{o,z,j}\mathbf{Z}_i + \lambda_j\xi_{i,d(j)} + \varepsilon_{ij} && i = 1,\ldots,n;\ j = 1,\ldots,p \\ \xi_{ik} &= \beta_{\text{D},\eta,k}\eta_i + \boldsymbol{\beta}_{\text{D},z,k}\mathbf{Z}_i + \phi_i + \psi_{ik} && k = 1,\ldots,d \end{aligned} \tag{2.3}$$

where $\boldsymbol{\beta}_{o,z,j}$ and $\boldsymbol{\beta}_{\text{D},z,k}$ are the $j$th and $k$th rows of the matrices $\boldsymbol{\beta}_{o,z}$ and $\boldsymbol{\beta}_{\text{D},z}$, respectively. Apply one more simplification, setting $\lambda_j = 1$ for all $j$ which yields

$$Y_{ij} = \alpha_j + (\beta_{\text{D},\eta,d(j)} + \beta_{o,\eta,j})\eta_i + (\boldsymbol{\beta}_{\text{D},z,d(j)} + \boldsymbol{\beta}_{o,z,j})\mathbf{Z}_i + \phi_i + \psi_{i,d(j)} + \varepsilon_{ij}. \tag{2.4}$$

To obtain (2.1) drop the intercept term $\alpha_j$ and use the random effect specification $\beta_{\text{D},\eta,k} \overset{\text{ind}}{\sim} N(b_\eta, \tau^2_{\text{D}})$, $\beta_{\text{D},z,k,\ell} \overset{\text{ind}}{\sim} N(b_{z,\ell}, \tau^2_{\text{D},\ell})$ $\beta_{o,\eta,j} \overset{\text{ind}}{\sim} N(0, \tau^2_o)$, and $\beta_{o,z,j,\ell} \overset{\text{ind}}{\sim} N(0, \tau^2_{o,\ell})$ for each $k$, $j$ and $\ell$, after standardizing the outcome variables, the covariates, and the exposure.

We will investigate models that fit into the framework (2.3), using the random effect assumption that $\beta_{o,\eta,j} \overset{\text{ind}}{\sim} N(0, \tau^2_o)$ and $\beta_{o,z,j,\ell} \overset{\text{ind}}{\sim} N(0, \tau^2_{o,\ell})$ and taking $\lambda_j = 1$ for the first outcome in each

domain. We will see that this framework is identifiable without further restrictions. In deriving model (2.3) we have made assumptions regarding the structure of the matrices $\mathbf{\Lambda}$ and $\mathbf{B}$; contrast these choices with those standard in the SEM literature, where it is conventional to assign the latent factors particular interpretations, such as "motor function" and "verbally mediated function." Having done that, one manually selects a small number of nonzero elements in the matrices $\mathbf{B}$ and $\mathbf{\Lambda}$ corresponding to hypothesized associations among the interpreted factors and between the interpreted factors and the individual outcome variables (Sanchez et al. 2005; Palomo, Dunson and Bollen 2007). For instance, Palomo et al. (2007) have a set of four measurements of democratization, measured separately for each country in 1960 and 1965. The four measurements from 1960 are taken to be indirect measurements of a latent factor capturing overall democratization in 1960, and similarly for 1965. Due to the temporal relationship, they assume that the latent democratization in 1960 had an effect on that in 1965 but not the other way around.

Like these authors, we associate each latent factor with a domain, and assign each of the outcomes to a single domain. However, we avoid manually specifying the relationships between the latent factors, instead making the assumption that the latent factors are related to each other by inclusion of the subject random effect $\phi_i$, and potentially by random effect modeling of the domain-specific coefficients $\beta_{\mathrm{D},\eta,k}$ and $\beta_{\mathrm{D},z,k,\ell}$.

The subject random effect $\phi_i$ captures positive correlation between all of the outcomes, after accounting for the exposure and covariates; to enforce this we restrict $\lambda_j > 0$ for each $j$. This is appropriate in many contexts, for instance the phthalates context of Section 4, the democratization example of Palomo et al. (2007), and the study by Thurston et al. (2009) of the effect of prenatal methylmercury exposure on neurodevelopmental outcomes. In the case where not all of the outcomes are positively correlated it may be possible to multiply some of the outcomes by $-1$ so that our models can be applied; for instance, in the methylmercury analysis of Budtz-Jorgensen et al. (2003), for most of the outcome variables a higher value indicates better neurological development, but in a few it indicates worse development. After multiplying the latter outcomes

by $-1$ our model should be applicable, since the outcomes would then presumably be positively correlated even after accounting for covariates and methylmercury exposure.

One could potentially also allow the case $\lambda_j = 0$ in addition to $\lambda_j > 0$. However, allowing $\lambda_j = 0$ simultaneously for all $j$ leads to non-identifiability (see Section 2.1). Additionally, it seems reasonable to require the latent factor associated with a domain to influence all of the outcomes within that domain. For these reasons we restrict $\lambda_j > 0$.

## 2.1 A Spectrum of Models

Next we define a set of models for grouped outcomes that span the spectrum from flexible latent factor models to parsimonious random effect models, and show conditions for identifiability. All variables are standardized before model-fitting.

**Model A**: Given in (2.3), treating $\beta_{\text{D},\eta,k}$, $\boldsymbol{\beta}_{\text{D},z,k}$, $\alpha_j$, and $\lambda_j$ as fixed effects restricting to $\lambda_j > 0$, and recalling that $\lambda_j = 1$ for the first outcome in each domain and that $\beta_{o,\eta,j} \sim N(0, \tau_o^2)$, $\beta_{o,z,j,\ell} \sim N(0, \tau_{o,\ell}^2)$, $\phi_i \sim N(0, \tau_\phi^2)$, and $\psi_{ik} \sim N(0, \tau_{\psi,k}^2)$ are random effects.

**Model B**: Identical to Model A except that it models $\beta_{\text{D},\eta,k}$ as a random effect, $\beta_{\text{D},\eta,k} \sim N(b_\eta, \tau_{\text{D}}^2)$ (reducing the effective number of free parameters in the model). Since this induces shrinkage of $\beta_{\text{D},\eta,k}$ across domains $k$, it is only reasonable if the effect of $\eta_i$ is believed to be similar for the outcomes in all domains.

**Model C**: Identical to Model A except that it sets $\boldsymbol{\beta}_{o,\eta} = \boldsymbol{\beta}_{o,z} = 0$.

**Model D**: Identical to Model C except that it additionally sets $\lambda_j = 1$ for every $j$.

**Model E**: Identical to Model D except that it additionally sets $\psi_{i,k} = \phi_i = 0$ for each $i$ and $k$, yielding the most parsimonious model:

$$Y_{ij} = \alpha_j + \beta_{\text{D},\eta,d(j)}\eta_i + \boldsymbol{\beta}_{\text{D},z,d(j)}\mathbf{Z}_i + \varepsilon_{ij}.$$

Model E is simply a regression model where the same coefficient is used for all outcomes within the same domain. It can be fit independently for each domain. Models A-E all include intercept terms $\alpha_j$ (unlike (2.1)), despite the fact that all variables are standardized. Dropping the intercept term would ignore the uncertainty in that intercept when estimating the parameters of interest, potentially leading to poor interval estimates.

Models A-D capture positive correlation between the outcomes, even after accounting for exposure and covariates. Models A-C allow the exposure effect $(\beta_{o,\eta,j} + \lambda_j \beta_{\mathrm{D},\eta,d(j)})$ and covariate effects $(\boldsymbol{\beta}_{o,z,j} + \lambda_j \boldsymbol{\beta}_{\mathrm{D},z,d(j)})$ to be different for outcomes within a domain. Model B allows shrinkage of the exposure effect across domains. In Appendix B we prove that our most general model (Model A) is identifiable so long as there is more than one domain and more than one outcome in each domain. If a particular domain has only one outcome, we show that identifiability can be achieved by setting $\beta_{o,\eta,j} = \tau^2_{\psi,d(j)} = 0 = \boldsymbol{\beta}_{o,z,j}$ for that outcome $j$.

We will use Bayesian inference in the above models. All prior distributions not given in the model descriptions are specified as follows. The parameters $\lambda_j$, $\alpha_j$, $\beta_{\mathrm{D},\eta,k}$, and the elements of the vector $\boldsymbol{\beta}_{\mathrm{D},z,k}$, are given prior distributions that are uniform on the real line, in the case of $\lambda_j$ restricting to positive values as discussed previously. For the variance parameters $\sigma^2_j$, $\tau^2_o$, $\tau^2_{o,\ell}$, $\tau^2_{\mathrm{D}}$, $\tau^2_\phi$, and $\tau^2_{\psi,k}$, we use a uniform prior on the associated standard deviation (Gelman, Carlin, Stern and Rubin 2004), with support on the interval from zero to two. This upper bound is reasonable: none of the variance parameters is expected to be greater than one due to the standardization of the outcomes, but we use a slightly higher upper bound since in some cases the likelihood can be high near and just above one. To understand why variance parameters with high likelihood typically have values $\leq 1$, take the example of Model A, and consider an arbitrary outcome $j$. Then

$$\mathrm{Var}(Y_{ij} | \alpha_j, \beta_{o,\eta,j}, \boldsymbol{\beta}_{o,z,j}, \beta_{\mathrm{D},\eta,d(j)}, \boldsymbol{\beta}_{\mathrm{D},z,d(j)}, \eta_i, \mathbf{Z}_i) = \lambda_j^2 \tau^2_\phi + \lambda_j^2 \tau^2_{\psi,d(j)} + \sigma^2_j. \tag{2.5}$$

For parameter vectors with high likelihood, the model typically explains some of the variability in the outcome $Y_{ij}$ in the sense that the conditional variance in (2.5) is $\leq \mathrm{Var}(Y_{ij})$. Due to standardization of $Y_{ij}$, we then have $\lambda_j^2 \tau^2_\phi + \lambda_j^2 \tau^2_{\psi,d(j)} + \sigma^2_j \leq 1$ for each $j$. This implies that $\sigma^2_j \leq 1$, and

taking $j$ to be the first outcome in an arbitrary domain $k$ yields $\lambda_j = 1$ and thus $\tau_\phi^2, \tau_{\psi,k}^2 \leq 1$. Finally consider $\tau_o^2$ and $\tau_{o,\ell}^2$, the variance of the regression coefficients $\beta_{o,\eta,j}$ and $\beta_{o,z,j,\ell}$. The values $(\beta_{o,\eta,j} + \lambda_j \beta_{\text{D},\eta,d(j)})$ are the correlations between $\eta_i$ and $Y_{ij}$ (which cannot be bigger than one in absolute value), so it is reasonable to expect the standard deviation $\tau_o$ of $\beta_{o,\eta,j}$ to be below one, and similarly for $\tau_{o,\ell}$.

While the prior distributions for $\lambda_j$, $\alpha_j$, $\beta_{\text{D},\eta,k}$, and $\boldsymbol{\beta}_{\text{D},z,k}$ are nonintegrable, the posterior distributions for Models A-E are integrable (well-defined). We assessed prior sensitivity for the variance parameters by changing the upper bounds on the standard deviations; the inferences reported in Sections 3-4 were insensitive to increases in the upper bound and to moderate decreases in the upper bound. Not surprisingly, decreasing the upper bound so far as to truncate the region of high likelihood changed the parameter estimates in an undesirable way.

### 2.2 COMPUTATION

Inference in Models A-E is performed by Markov chain Monte Carlo. We verify convergence of the Markov chain to the posterior distribution using the Gelman-Rubin diagnostic (checking that the scale reduction factor is less than 1.2 for each parameter; Gelman and Rubin 1992), and ensuring that the Monte Carlo standard error as estimated using consistent batch means is less than 0.5% of each parameter's point estimate (Flegal, Haran and Jones 2008). For the simulation study we allow slightly higher Monte Carlo standard error, roughly 2%, to reduce computation time since we analyze a large number of treatment combinations. Parameter point estimates are taken to be the posterior mean, and $(1-a)$ interval estimates for $a \in (0,1)$ are given by the $a/2$ and $(1-a/2)$ quantiles of the posterior distribution. For the variance parameters and $\lambda_j$ we find the posterior mean on the log scale and then exponentiate to obtain the point estimate, since the posterior distributions of these parameters are right-skewed.

# 3 SIMULATION STUDY

## 3.1 EXPERIMENTAL CONDITIONS

We performed a simulation study to compare the performance of the five models and the trade-off between flexibility and parsimony. We used one exposure $\eta_i$ and one covariate $Z_i$, which were generated from a bivariate normal distribution with mean zero, variance one and covariance 0.2 and then standardized. The value of $\mathbf{Y}_i$ was then simulated according to the models, with the standard deviation (conditional on $\eta_i, Z_i$, and all model parameters) of each $Y_{ij}$ set to one, i.e. $\sigma_j = 1$. The outcomes were not standardized as described in Section 2, because they are already standardized in expectation.

Our primary comparison in the simulation study was between Models A and B or A and D. In additional results not reported here, the performance of Model C was generally in between those of Models D and A, and was similar to A when parameters $\beta_{o,\eta,j}$ and $\beta_{o,z,j}$ were not too large. We leave out Model E because it does not incorporate dependence across domains; however, Model E is explored in Section 4.

In addition to varying the models used to simulate and fit the data, we also varied six factors each with two levels: (a) sample size (n=500 or 250); (b) number of outcomes in each domain, either (4, 2, 1) or (4, 6, 3); (c) $\lambda_j$ values (all equal to 1, or some $\neq 1$); (d) $\bar{\beta}_{\text{D},\eta}$, defined as $\frac{1}{3}\sum_{k=1}^{3}\beta_{\text{D},\eta,k}$, either (0.05 or 0.15); (e) $\beta_{\text{D},\eta,k} - \bar{\beta}_{\text{D},\eta}$, either (-0.05, 0, 0.05) or (-0.1, 0, 0.1), and (f) $\beta_{\text{D},z,k} - \bar{\beta}_{\text{D},z}$, either (-0.05, 0, 0.05) or (-0.2, 0, 0.2) where $\bar{\beta}_{\text{D},z}$ is defined as $\frac{1}{3}\sum_{k=1}^{3}\beta_{\text{D},z,k}$. For simulations in which some $\lambda_j \neq 1$, for the model with (4,2,1) outcomes we used $\lambda_j$ values of $(1, 1, 0.5, 0.5,\ \ 1, 0.5,\ \ 1)$ and for the model with (4,6,3) outcomes, we used $\lambda_j$ values of $(1, 1, 0.5, 0.5,\ \ 1, 1, 1, 0.5, 0.5, 0.5,\ \ 1, 0.5, 0.5)$. For all models we used $\bar{\beta}_{\text{D},z} = 0.2$, $\tau_\phi = 0.2$, and $\tau_{\psi,k} = 0.05$ for $k = 1, 2, 3$. For Models A and B, the standard deviations $\tau_o$ and $\tau_{o,1}$ of $\beta_{o,\eta,j}$ and $\beta_{o,z,j}$ were set to 0.05.

For each treatment combination, 25 datasets were generated. Instead of performing a full

factorial design we analyze a subset of treatment combinations, since a full factorial design requires fitting the models exponentially many times and since Markov chain methods require nontrivial computational time. In all estimation models, the value of $\lambda_j$ for the first outcome in each domain was set to one as discussed in Section 2. Convergence was diagnosed as discussed in Section 2.2.

### 3.2 EVALUATION

We evaluated the bias and the root mean squared error (RMSE) for ten parameters of interest, such as $\beta_{o,\eta,j}$ and $\sigma_j^2$. Since each model has multiple $\sigma_j^2$ parameters, for instance, we averaged the bias and RMSE across the multiple parameters in such cases. Table 1 shows results for a variety of conditions when data are simulated and analyzed under Model A. Table 2 gives results from a small number of conditions when Models A, B, or D are the true model and the data are analyzed under the same or different models. The final rows are for the outcome-specific slopes $\beta_{o,\eta,j} + \lambda_j\ \beta_{\mathrm{D},\eta,d(j)}$, labeled 'OS $\eta$', and $\beta_{o,z,j} + \lambda_j \beta_{\mathrm{D},z,d(j)}$, labeled 'OS $z$', which are the parameters of primary interest. The bias of $\lambda_j$ is presented on the log scale because the distribution of $\lambda_j$ estimates was very right-skewed (a few of the simulated datasets have several $\lambda_j$ estimates greater than 10).

These tables show that the parameter biases and RMSE are small for most parameters, and in particular for the parameters of interest OS $\eta$ and OS $z$. These parameters are estimated with little bias and low RMSE under all conditions. The true values of the parameters OS $\eta$ in this simulation study typically range from $-.15$ to .35, while the values of the OS $z$ parameters range from $-.1$ to .5. By contrast the absolute bias for all these parameters is below .015 under all conditions in the Tables, and the RMSE is always less than .08, almost an order of magnitude smaller than the parameter range. The bias and RMSE are even smaller if we look only at the conditions having the larger sample size $n = 500$ and for which the fitted model is at least as general as the simulated model (the latter eliminates Columns 3, 5, and 6 of Table 2). In these cases the absolute bias is less than .008, and the RMSE is less than .05.

As expected the smaller sample size $n = 250$ gives higher RMSE for OS $\eta$ and OS $z$ than $n = 500$; compare Column 2 to Column 3 and Column 4 to Column 5 in Table 1. Also, the RMSE for these parameters is higher when the simulation model is A and we fit the simpler model D; compare Column 2 to Column 3 and Column 4 to Column 6 in Table 2. This is not surprising since the simulated values for OS $\eta$ and OS $z$ are different for outcomes in the same domain, and Model D restricts these to have a common value within a domain. On the other hand, when the simulated model is D the RMSE of OS $\eta$ and OS $z$ are smaller when fitting Model D than when fitting Model A; compare Columns 9 and 10 in Table 2. This demonstrates the advantage of fitting a simpler model when that model is correct. Interestingly, we do not see this advantage when the simulated model is B; compare Columns 7 and 8 in Table 2. In this case fitting Model A gives nearly the same bias and RMSE as fitting Model B, for almost all the parameters including OS $\eta$ and OS $z$. In general the results from fitting Model B are very similar to those from fitting Model A (see also Columns 4-5 of Table 2), and we do not find an advantage to using Model B over Model A.

Although the estimates of the outcome-specific effects $(\text{OS } \eta)_j = \beta_{o,\eta,j} + \lambda_j \ \beta_{\text{D},\eta,d(j)}$ and $(\text{OS } z)_j = \beta_{o,z,j} + \lambda_j \beta_{\text{D},z,d(j)}$ are very accurate under all experimental conditions, the estimates of the individual components $\lambda_j$ and $\beta_{\text{D},\eta,k}$, $\beta_{\text{D},z,k}$ are sometimes less accurate. In particular, when both simulating from and fitting Model A, with $(4,6,3)$ outcomes per domain and the smaller sample size $n = 250$ (Columns 6-10 of Table 1), $\lambda_j$ is overestimated while $\beta_{\text{D},\eta,k}$ and $\beta_{\text{D},z,k}$ are underestimated. However, this is greatly alleviated by increasing the sample size; compare the last column of Table 1 to the first column of Table 2, which are identical treatment conditions except that the latter has the larger sample size $n = 500$. The bias ($\times 100$) of $\log \lambda_j$ drops from 111.56 to 16.52, and similarly for $\beta_{\text{D},\eta,k}$ and $\beta_{\text{D},z,k}$. So caution should be taken in interpreting the $\lambda_j$ estimates when the sample size is small, but the estimates of the outcome-specific effects OS $\eta$ and OS $z$ are reliable.

The residual standard deviations $\sigma_j$ are also estimated well under all conditions. The true value

is $\sigma_j = 1$, while the absolute bias and the RMSE are less than .02 and .08, respectively (Tables 1-2). Not surprisingly the standard deviation $\tau_{\psi,k}$ of the domain-specific random effects is estimated less accurately when there are fewer outcomes within each domain; see Table 1. For all simulation conditions $\tau_{\psi,k}$'s are biased upwards, and $\tau_\phi$ is biased downwards. This means that more of the individual variation between outcomes is attributed to the domain-specific effects than is actually the case. However this bias is smaller in simulations with many outcomes and a larger sample size; compare Table 1 to Table 2.

In summary, this simulation study evaluated the accuracy of point estimation. It showed that the parameters of most interest, the outcome-specific exposure and covariate effects, are estimated well under all models and simulation conditions. They are estimated even more accurately when restricting to the larger sample size, and when the fitted model is at least as general as the simulation model. We do not find an advantage to fitting Model B over Model A under any of the experimental conditions. Also, there is a deleterious effect of fitting the simpler Model D when the data come from Model A, in terms of accuracy of the outcome-specific effect estimates. When Model D is the truth, there is an additional improvement in accuracy obtained by fitting Model D, but the estimates from Model A are still very accurate. Due to these facts, this simulation study suggests that Model A is an excellent general-purpose choice from the perspective of accurate point estimation, and that Model D has even better statistical efficiency if the data do indeed come from such a parsimonious model. In addition to point estimation one should consider the width of interval estimates and the power to detect exposure and covariate effects, which we explore in Section 4.

# 4 PHTHALATES ANALYSIS

## 4.1 DATA

Phthalates occur in a variety of industrial and household products including cosmetics, children's toys, and baby care products and use of some of these products has been linked to elevated levels of phthalates in humans (Hauser and Calafat 2005; Sathyanarayana et al. 2008). The chemicals are believed to affect reproductive hormone concentration and anogenital distance in male infants, via an anti-androgenic and possibly estrogenic mechanism (Gray et al. 2006). This leads to the hypothesis that they may affect other sexually dimorphic traits in male infants, including anthropometric measures like head circumference, weight and skinfold thickness.

The Study for Future Families is a multicenter pregnancy cohort study relating maternal phthalate levels to a variety of infant anthropometric and reproductive characteristics (Swan et al. 2003). The data include measurements of phthalate metabolite concentrations in maternal urine, infant anthropometry measurements, and relevant covariate information for several hundred pregnancies. Measurements of eight distinct phthalate metabolites are available, most of which are highly correlated. We will summarize these measurements and relate them to the anthopometry outcomes via the phthalate "score." This score (defined in Swan 2008) is a summary of the five phthalate measurements that have been found to be related to anogenital distance in male infants; this score falls in the range $(0-15)$, with 15 representing the highest exposure. One can instead directly use all the phthalate measurements by taking $\eta_i$ to be a latent variable that represents overall phthalate exposure, and adapting Models A-E to this context (in the manner of a structural equation model). Explicitly, this approach assumes that $\eta_i \sim N(0,1)$ and models the multiple phthalate measurements $X_{i\ell}$ indexed by $\ell$ as $X_{i\ell} \overset{\text{ind}}{\sim} N(\alpha_{X\ell} + \lambda_{X\ell}\eta_i, \delta_\ell^2)$. Our results from this approach were qualitatively very similar to those reported here.

The infant anthropometry measurements include four skinfold thickness metrics plus body mass index, weight percentile-for-age, and head circumference percentile-for-age. Most of these

traits are known to be strongly sexually dimorphic; in particular, skinfold thickness measures tend to be larger in females at all ages up to three years, while head circumference and weight are larger in males than females of the same age (Rodriguez et al. 2004; U.S. Centers for Disease Control and Prevention 2000). The case of body mass index (BMI) is a bit more subtle, since although it tends to be slightly larger in boys then girls at birth, this may change with age; additionally, BMI is not typically used as a clinical measure of body fatness or composition in infants (Wells 2000; Brock, Falcao and Leone 2008). We include it here for completeness. The anthropometry metrics fall into three natural domains, namely (1) the skinfold thickness metrics, (2) weight percentile and BMI, which are closely related, and (3) head circumference percentile. We address the hypothesis that, due to an anti-androgenic mechanism, phthalates affect anthropometry measures in male infants, causing these measures to be more like those of females. For the above outcomes that means higher skinfold thickness, lower weight, and smaller head circumference.

Covariates are available including infant's age and gestational age, mother's age at time of birth, mother's race, mother's educational level, mother's smoking status, and the creatinine concentration for the urine sample from which the phthalate measurements were taken. We performed a preliminary analysis by regressing each of the anthropometry measurements on the covariates and phthalate score. In this and the rest of our regression analyses, we used the following transformations of the variables in order to make the assumptions of the linear models most reasonable: a square root transformation of the skinfold thickness measurements and creatinine concentration; a logistic transformation for the weight percentile and head circumference percentile, after rounding the smallest measurements up to 0.001 and the largest down to .999; and a log transformation of the phthalate score. After these transformations, all variables are standardized before fitting regression models, as described in Section 2.1.

Fitting the separate linear regression models for the anthropometry outcomes and applying backward elimination, we found strong evidence of a relationship between infant's age, gestational age, mother's age, and mother's race (categorizing into caucasian / non-) and some of the

anthropometry measures. We include these four predictors in our models, as well as the creatinine concentration. Although the latter was not found to be a significant predictor for any of the anthropometric outcomes, it is a significant predictor for the phthalate concentrations and is included in order to adjust for this effect (Barr et al. 2005). We restrict to infants for which complete data are available for the phthalate and anthropometry measurements, age, gestational age, mother's age, mother's race, and creatinine, leaving 118 male infants out of 172.

Table 3 gives summary statistics for the anthropometry metrics and covariates, and relates these variables to the phthalate score. It shows the average phthalate score for individuals in each category (for categorical variables) or individuals above and below the variable median (for continuous variables). It also shows the regression coefficient for phthalate score obtained from the separate regressions on standardized data described above, along with 95% confidence intervals. In Table 3 the phthalate score is empirically higher for Caucasian mothers, younger mothers, in cases of younger infants or infants with a lower gestational age, and in cases with higher creatinine concentrations. We do not see a significant relationship between phthalate score and the outcome variables using the separate regressions; the hope is that by fitting the multiple-outcomes models from Section 2.1 we will have higher power to detect a phthalate effect if one exists. Some but not all of the point estimates of the phthalate coefficients in Table 3 are consistent with the hypothesized effect.

### 4.2 RESULTS

Next we apply Models A, C, D, and E to the phthalates data. We omit Model B from consideration because we do not hypothesize that the phthalate effect is similar across domains (e.g., it is not hypothesized to have the same sign in all domains). The third domain defined in Section 4.1 has only one outcome variable (head circumference), so for identifiability in Models A,C,D,E we set $\beta_{o,\eta,j} = \tau^2_{\psi,d(j)} = 0 = \boldsymbol{\beta}_{o,z,j}$ for this outcome $j$. To see whether Models A, C, and D were appropriate we verified that the residuals from the separate regressions done in Section 4.1 were

positively correlated for the different outcomes, which they were (having correlations $.08 - .65$).

The estimates of the phthalate effects $(\text{OS } \eta)_j = (\beta_{o,\eta,j} + \lambda_j \beta_{\text{b},\eta,d(j)})$ for each of the outcome variables $j$ and each of the models A,C,D,E are shown in Table 4. While the point estimates are similar for Models D and E, the posterior 95% intervals are wider for Model D, reflecting the fact that Model E is more parsimonious. However, there is evidence that Model D fits the data better: the random effects $\phi_i$ and $\psi_{i,k}$, which are missing in Model E, have high variance in Model D (the point and 95% interval estimates of $\tau_\phi$, $\tau_{\psi,1}$, and $\tau_{\psi,2}$ are .50 $(.40, .61)$, .22 $(.025, .41)$, and .54 $(.38, .70)$, respectively).

The width of the posterior intervals decreases from left to right in Table 4 as the parsimony of the model increases; the average interval width in the Table is 47.4, 44.5, 38.9, and 27.1 for Models A, C, D, and E respectively. Although Models A, C, D, and E are more parsimonious than separate regression models, we still do not find a significant relationship between phthalate exposure and the anthropometry outcomes in any of the models. While links have been found between phthalate exposure and other sexually dimorphic traits, such a link may not exist between phthalate exposure and the anthropometry measures investigated here. Alternatively, such a link may exist and we may not be able to detect it due to the limited size of our dataset and the need to adjust for a number of covariates, including the creatinine concentration and infant's age. There is also some inflation of the uncertainty in the phthalate score regression coefficients, due to the correlation between phthalate score and creatinine. In a standard multiple linear regression model the magnitude of this effect could be measured using the variance inflation factor (Neter, Kutner, Nachtsheim and Wasserman 1996). Although it is not obvious how to calculate a variance inflation factor for our multiple-outcomes models, in a standard linear regression model with our set of covariates the variance inflation factor for phthalate score would be 2.44, which is low, indicating that this effect is not a source of concern.

Since we have not detected a phthalate effect using any of the models, we will demonstrate the differences between Models A,C,D,E using differences in the covariate effects. The covariate

effects $(\text{OS}\ \mathbf{Z})_j = \boldsymbol{\beta}_{o,z,j} + \lambda_j \boldsymbol{\beta}_{\text{D},z,d(j)}$ for the simplest model, Model E, are shown in Table 5. In this model the covariate effects are domain-specific but not outcome-specific ($(\text{OS}\ \mathbf{Z})_j$ simplifies to $\boldsymbol{\beta}_{\text{D},z,d(j)}$), so we display them by domain. We find a positive relationship between gestational age and the BMI and weight percentile outcomes. This is in accordance with previous findings of a positive correlation between gestational age and BMI/weight (U.S. Centers for Disease Control and Prevention, National Center for Health Statistics 2000). Some drawbacks of the CDC weight-for-age growth chart in relation to the WHO growth charts have also been noted (de Onis, Garza, Onyango and Borghi 2007). We also find a negative relationship between mother's age and the skinfold thickness outcomes, which is plausible although previously unobserved. We find a negative relationship between infant age and the outcomes in domains one and two, namely the skinfold thickness, BMI, and weight percentile outcomes. Such relationships are surprising, in part since weight percentile is already adjusted for infant age. These negative correlations also exist in the raw data; however, if we restrict to infants in the most rapid phase of growth (younger than 9 months) these become positive correlations, which are in accordance with previous findings.

The point estimates for the parameters $(\text{OS}\ \mathbf{Z})_j$ in Model D are nearly identical to those for Model E as reported in Table 5 (on average differing by only 2.1%). However, the 95% intervals for $(\text{OS}\ \mathbf{Z})_j$ are substantially wider in Model D than in Model E (wider on average by 46%). Because of this, the effect of mother's age in Domain 1 is not significant in Model D. This illustrates the fact that Model E is more parsimonious than Model D.

Models A and C allow the covariate effects $(\text{OS}\ \mathbf{Z})_j$ to be different for each outcome in a domain, while Models D and E do not. For the phthalates data this model flexibility comes with a price in terms of the width of the 95% posterior intervals for $(\text{OS}\ \mathbf{Z})_j$. The average width of these intervals is 34.8, 27.5, 26.6, and 18.2 (strictly decreasing) for Models A, C, D, and E respectively. However, Model C has the same number of significant covariate effects (8 significant effects) as Model D for the phthalates data, and Model A has one more (9 effects). Model C has the same set of significant regression coefficients as Model D, while Model A has a slightly different set. The

following effects are significant in Models C and D but not in Model A (listing covariate / outcome pairs) : (1) infant age / skinfold thickness quadriceps; (2) infant age / BMI; (3) gestational age / BMI. The following effects are significant in Model A but not in Model C: (1) mother's age / skinfold thickness quadriceps; (2) mother's age / skinfold thickness triceps; (3) gestational age / skinfold thickness triceps; (4) race / skinfold thickness subscapular.

Given the discrepancy between the conclusions that would be drawn from Models A and C, we attempt to determine which model is more believable in the phthalates context. We do this by finding evidence to corroborate the coefficient point estimates from one of these two models. When the point estimates from Model A are averaged over the outcomes in each domain, the results are very close to the estimates of $(\text{OS } \mathbf{Z})_j$ from Models D and E (on average differing from Model D by 3.1% and from Model E by 2.9%). However, when the same procedure is done for Model C, the estimates differ substantially from Models D and E (on average differing from Model D by 47% and from Model E by 46%).

We can additionally compare the results from Models A and C to the separate regressions that were fit to each outcome variable in Section 4.1. Since Models A and C fit a joint model to all of the outcome variables, we can expect the point estimates of $(\text{OS } \mathbf{Z})_j$ for outcomes in the same domain to be shrunk towards each other, relative to the point estimates from separate regression models. So we don't expect the point estimates from Models A and C to be very close to those from the separate regressions, but we might expect the point estimates *averaged over a domain* to be similar between the separate regressions and Models A and C. Indeed, the point estimates from from Model A, averaged over each domain, are very similar to those from the separate regressions (differing on average by only 5.0%). However, the point estimates from Model C, averaged over each domain, are very dissimilar to those from the separate regressions (differing on average by 62%). So we can find little corroboration for the point estimates obtained by Model C, and find the results from Model A to be more believable.

The relationship of the estimated effects $(\text{OS } \mathbf{Z})_j$ from Models A, C, D, E, and those obtained

from separate regressions are illustrated in Figure 1. Estimates from Model A show shrinkage towards the domain average, relative to estimates from the separate regressions. Models D and E restrict to a single coefficient estimate per domain, and this estimate is close to the average of the outcome-specific estimates from Model A and close to the average of the outcome-specific estimates from separate regressions. By contrast, no shrinkage between domains is visible; this is due to the fact that the domain-specific coefficients $\boldsymbol{\beta}_{\mathrm{D},z,k}$ are fixed effects in these models. The coefficient estimates from Model C, shown at the bottom of Figure 1, are substantially different from those of the other models.

The shrinkage of coefficients in Model A relative to separate regressions appears moderate in Figure 1. However, in cases where the signal-to-noise ratio is low this shrinkage can be dramatic. For the phthalates data this can be seen by comparing Table 3, Column 4 to Table 4, Column 1. The coefficient estimates are strongly pulled together. Due to the ability of Model A to pool information across outcomes, it detects one more significant covariate effect than do the separate regression models (9 vs. 8), namely between the covariate "infant age" and the outcome "skinfold thickness triceps."

# 5 CONCLUSIONS

We introduced models for regression with multiple outcomes nested in domains, that span the spectrum from very general continuous latent factor models to very parsimonious random effect models. These extend the model of Thurston et al. (2009) in the sense that some of them allow outcome-specific weights $\lambda_j$ for the latent factors associated with the domains. We evaluated these models on simulated data and on the phthalates data. The simulation study found the most general model, Model A, to be very accurate in terms of point estimation of the outcome-specific exposure and covariate effects, regardless of which of the models is used to simulate the data. It also found an additional advantage to fitting the simpler Model D when the data are drawn from Model D, but

no advantage to Model B under any of the experimental conditions. In the models that incorporate the $\lambda_j$ parameters, they are sometimes estimated with substantial upward bias, but this effect is alleviated by increasing the sample size, and the bias does not affect the accuracy for the outcome-specific effects of interest.

In the phthalates analysis Model B was not appropriate and Models A, C, D, and E were considered. None of the models found a phthalate effect on any of the outcomes, so we focused on the covariate effects to investigate the differences between the models. We found that Models A, D, and E gave point estimates that were reasonable and consistent with one another, while those from Model C were distinct and uncorroborated by other evidence. This suggests that it is disadvantageous to incorporate the parameters $\lambda_j$ without also including outcome-specific random effects $\beta_{o,\eta,j}$ and $\boldsymbol{\beta}_{o,z,j}$. While the interval estimates were wider in Model A than in Model D, a larger number of significant covariate effects were found in Model A. Model A has the advantage that it is much more flexible than the other models, and easier to justify since it allows the exposure and covariate effects to differ across outcomes within a domain. Since there was no loss of the ability to detect covariate effects relative to Model D, Model A is the most reasonable choice for the phthalates data. Although Model E found more significant covariate effects than any other model, it is a very difficult model to justify for the phthalates data. This is in part because it assumes conditional independence of the outcomes given the covariates, exposure, and regression coefficients, a property which is clearly violated for these data.

In summary, we found that Models A and D performed better than the other three models for both simulated and real data. Model A is much more flexible than Model D in the sense of allowing the exposure and covariate effects to differ for outcomes in the same domain, so we recommend this model for general use. In some cases where the effects might plausibly be the same for outcomes within each domain (after standardization), and where statistical efficiency is of primary concern due to a small sample size, Model D is also a good choice.

# REFERENCES

Barr, D. B., Wilder, L. C., Caudill, S. P., Gonzalez, A. J., Needham, L. L., and Pirkle, J. L. (2005), "Urinary creatinine concentrations in the U.S. population: Implications for urinary biologic monitoring measurements," *Environmental Health Perspectives*, 113, 192–200.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, New York: Wiley & Sons.

Brock, R. S., Falcao, M. C., and Leone, C. (2008), "Body mass index values for newborns according to gestational age," *Nutrición Hospitalaria*, 23, 487–492.

Budtz-Jorgensen, E., Keiding, N., Grandjean, P., and Weihe, P. (2002), "Estimation of health effects of prenatal methylmercury exposure using structural equation models," *Environmental Health*, 1, 2.

Budtz-Jorgensen, E., Keiding, N., Grandjean, P., Weihe, P., and White, R. F. (2003), "Statistical methods for the evaluation of health effects of prenatal exposure," *Environmetrics*, 14, 105–120.

Coull, B. A., Hobert, J. P., Ryan, L. M., and Holmes, L. B. (2001), "Crossed random effect models for multiple outcomes in a study of teratogenesis," *Journal of the American Statistical Association*, 96, 1194–1204.

de Onis, M., Garza, C., Onyango, A., and Borghi, E. (2007), "Comparison of the WHO child growth standards and the CDC 2000 growth charts," *The Journal of nutrition*, 137(1), 144.

Dunson, D. B. (2000), "Bayesian latent variable models for clustered mixed outcomes," *Journal of the Royal Statistical Society, Series B.*, 62, 355–366.

Flegal, J. M., Haran, M., and Jones, G. L. (2008), "Markov chain Monte Carlo: Can we trust the third significant figure?," *Statistical Science*, 23, 250–260.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn, Boca Raton, FL: Chapman and Hall.

Gelman, A., and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Gray, L. E., Wilson, V. S., Stoker, T., Lambright, C., Furr, J., Noriega, N., Howdeshell, K., Ankley, G. T., and Guillette, L. (2006), "Adverse effects of environmental antiandrogens and androgens on reproductive development in mammals," *International Journal of Andrology*, 29, 96–104.

Hauser, R., and Calafat, A. M. (2005), "Phthalates and human health," *Occupational and Environmental Medicine*, 62, 806–818.

Lin, X., Ryan, L., Sammel, M., Zhang, D., Padungtod, C., and Xu, X. (2000), "A scaled linear mixed model for multiple outcomes," *Biometrics*, 56, 593–601.

Muthén, B. (2002), "Beyond SEM: General latent variable modeling," *Behaviormetrika*, 29, 81–117.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models*, 4th edn, Boston: McGraw-Hill.

Palomo, J., Dunson, D. B., and Bollen, K. (2007), Bayesian structural equation modeling,, in *Handbook of Latent Variable and Related Models*, ed. S. Lee, Elsevier, Amsterdam, pp. 163–188.

Rodriguez, G., Samper, M. P., Ventura, P., Moreno, L. A., Olivares, J. L., and Perez-Gonzalez, J. M. (2004), "Gender differences in newborn subcutaneous fat distribution," *European Journal of Pediatrics*, 163, 457–461.

Roy, J., Lin, X., and Ryan, L. M. (2003), "Scaled marginal models for multiple continuous outcomes," *Biostatistics*, 4, 371–383.

Sammel, M. D., and Ryan, L. M. (1996), "Latent variable models with fixed effects," *Biometrics*, 52, 650–663.

Sammel, M., Lin, X., and Ryan, L. (1999), "Multivariate linear mixed models for multiple outcomes," *Statistics in Medicine*, 18, 2479–2492.

Sanchez, B. N., Budtz-Jorgensen, E., Ryan, L. M., and Hu, H. (2005), "Structural equation models: A review with applications to environmental epidemiology," *Journal of the American Statistical Association*, 100, 1443–1455.

Sathyanarayana, S., Calafat, A. M., Liu, F., and Swan, S. H. (2008), "Maternal and infant urinary phthalate metabolite concentrations: Are they related?," *Environmental Research*, 108, 413–418.

Sathyanarayana, S., Karr, C. J., Lozano, P., Brown, E., Calafat, A. M., Liu, F., and Swan, S. H. (2008), "Baby care products: Possible sources of infant phthalate exposure," *Pediatrics*, 121, 260–268.

Swan, S. H. (2008), "Environmental phthalate exposure in relation to reproductive outcomes and other health endpoints in humans," *Environmental Research*, 108, 177–184.

Swan, S. H., Brazil, C., Drobnis, E. Z., Liu, F., Kruse, R. L., Hatch, M., Redmon, J. B., Wang, C., and Overstreet, J. W. (2003), "Geographic differences in semen quality of fertile U. S. males," *Environmental Health Perspectives*, 111, 414–420.

Swan, S. H., Main, K. M., Liu, F., Stewart, S. L., Kruse, R. L., Calafat, A. M., Mao, C. S., Redmon, J. B., Ternand, C. L., Sullivan, S., and Teague, J. L. (2005), "Decrease in anogenital distance among male infants with prenatal phthalate exposure," *Environmental Health Perspectives*, 113, 1056–1061.

Thurston, S. W., Ruppert, D., and Davidson, P. W. (2009), "Bayesian models for multiple outcomes nested in domains," *Biometrics*, 65, 1078–1086.

U.S. Centers for Disease Control and Prevention, National Center for Health Statistics (2000), "CDC Growth Charts,". URL: http://www.cdc.gov/growthcharts.

Wells, J. C. (2000), "A Hattori chart analysis of body mass index in infants and children," *International Journal of Obesity and Related Metabolic Disorders*, 24, 325–329.

| Num of Outcomes | | 4,2,1 | | | | 4,6,3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\bar{\beta}_{D,\eta}$ | | 0.05 | | | | 0.05 | | 0.15 | | 0.15 |
| sd($\beta_{D,\eta}$), sd($\beta_{D,z}$) | | Small | | | | Small | Large | Small | Large | Large |
| $\lambda_j$ | | Same | | Different | | Same | | | | Diff |
| $n$ | | 250 | 500 | 250 | 500 | 250 | | | | |
| bias $\sigma_j$ | | -1.39 | -1.07 | -1.99 | -1.61 | -0.46 | -0.78 | -0.38 | -0.29 | -0.68 |
| bias $\tau_\phi$ | | -4.80 | -3.14 | -8.40 | -6.58 | -7.74 | -10.93 | -10.30 | -8.85 | -13.15 |
| bias $\tau_{\psi,k}$ | | 8.68 | 6.49 | 11.60 | 10.38 | 3.52 | 3.25 | 1.92 | 2.21 | 4.12 |
| bias $\log(\lambda_j)$ | | -22.89 | -20.63 | 8.14 | -13.38 | 46.97 | 72.99 | 59.70 | 45.92 | 111.56 |
| bias $\beta_{o,\eta,j}$ | $\times 100$ | 0.12 | 0.45 | -0.12 | 0.47 | 0.02 | 0.56 | 0.78 | 0.78 | 0.16 |
| bias $\beta_{o,z,j}$ | | 0.90 | 1.45 | 0.52 | 1.00 | 1.46 | 1.49 | 0.81 | 1.09 | 1.14 |
| bias $\beta_{\mathrm{D},\eta,k}$ | | 0.32 | 0.72 | 1.36 | -0.29 | -1.03 | -0.32 | -4.46 | -2.72 | -3.49 |
| bias $\beta_{\mathrm{D},z,k}$ | | -0.49 | -0.34 | -1.18 | -1.73 | -4.07 | -4.23 | -5.74 | -2.28 | -4.64 |
| bias OS $\eta$ | | 0.30 | 0.51 | 0.84 | 0.26 | -0.51 | 0.28 | -0.88 | -0.80 | -0.34 |
| bias OS $z$ | | -1.23 | -0.38 | -0.50 | -0.77 | -0.56 | -1.50 | -1.09 | 0.10 | -0.64 |
| RMSE $\sigma_j$ | | 5.59 | 4.33 | 7.24 | 6.24 | 5.20 | 5.30 | 5.22 | 5.33 | 5.68 |
| RMSE $\tau_\phi$ | | 7.51 | 6.30 | 9.26 | 7.84 | 10.83 | 12.74 | 11.05 | 10.85 | 14.75 |
| RMSE $\tau_{\psi,k}$ | | 9.86 | 7.78 | 13.60 | 12.28 | 7.80 | 7.63 | 5.60 | 5.45 | 11.98 |
| RMSE $\log(\lambda_j)$ | | 44.48 | 42.59 | 52.28 | 46.63 | 92.68 | 115.16 | 83.94 | 80.39 | 141.57 |
| RMSE $\beta_{o,\eta,j}$ | $\times 100$ | 3.87 | 4.02 | 3.72 | 3.76 | 4.23 | 4.70 | 4.90 | 4.57 | 4.74 |
| RMSE $\beta_{o,z,j}$ | | 4.24 | 4.49 | 4.45 | 4.01 | 5.79 | 5.61 | 5.18 | 5.23 | 5.37 |
| RMSE $\beta_{\mathrm{D},\eta,k}$ | | 6.51 | 4.46 | 5.97 | 5.15 | 4.26 | 4.61 | 6.88 | 6.38 | 7.58 |
| RMSE $\beta_{\mathrm{D},z,k}$ | | 5.58 | 4.67 | 7.24 | 5.70 | 8.79 | 9.05 | 8.47 | 7.39 | 9.53 |
| RMSE OS $\eta$ | | 5.56 | 4.10 | 5.19 | 3.77 | 5.02 | 5.07 | 5.19 | 5.27 | 5.74 |
| RMSE OS $z$ | | 5.81 | 4.16 | 5.71 | 4.39 | 6.03 | 5.52 | 5.84 | 5.55 | 5.62 |

Table 1. Values of bias$\times$100 and RMSE$\times$100, for some particular treatments in the simulation study. In all of these cases, we use data simulated from Model A and we report the results for fitting Model A. Also, recall the true values of $\tau_o = \tau_{o,1} = 0.05$, $\sigma_j = 1$, $\tau_\phi = 0.2$, and $\tau_{\psi,k} = 0.05$.

| $\lambda_j$ | | Different | | Same | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Simulated | | Model A | | Model A | | | Model B | | Model D | |
| Estimation | | A | D | A | B | D | A | B | A | D |
| bias $\sigma_j$ | | -0.36 | 0.68 | -0.30 | -0.30 | 0.15 | -0.01 | -0.00 | -0.28 | -0.09 |
| bias $\tau_\phi$ | | -7.03 | -7.89 | -3.91 | -3.92 | -1.44 | -2.92 | -3.13 | -3.44 | -1.41 |
| bias $\tau_{\psi,k}$ | | 4.08 | 1.27 | 2.66 | 2.64 | 2.73 | 3.25 | 2.99 | 2.75 | 2.84 |
| bias $\log(\lambda_j)$ | | 16.52 | 48.52 | -0.10 | -0.04 | 0.00 | -3.06 | -0.90 | 0.96 | 0.00 |
| bias $\beta_{o,\eta,j}$ | | 0.21 | -0.33 | 0.68 | 0.74 | -0.03 | 0.39 | 0.42 | 0.17 | 0.00 |
| bias $\beta_{o,z,j}$ | $\times 100$ | 0.96 | 0.43 | 0.44 | 0.45 | -0.06 | 0.39 | 0.45 | 0.24 | 0.00 |
| bias $\beta_{\mathrm{D},\eta,k}$ | | -1.08 | -4.12 | -1.35 | -1.69 | -0.06 | -0.22 | -0.54 | -0.80 | -0.01 |
| bias $\beta_{\mathrm{D},z,k}$ | | -2.18 | -6.59 | -0.61 | -0.90 | 0.34 | -0.40 | -0.75 | -1.49 | -0.25 |
| bias OS $\eta$ | | -0.05 | 0.18 | -0.52 | -0.53 | -0.28 | 0.07 | 0.04 | -0.39 | -0.13 |
| bias OS $z$ | | -0.01 | 0.16 | -0.02 | -0.02 | 0.16 | -0.28 | -0.28 | -0.29 | -0.09 |
| RMSE $\sigma_j$ | | 3.36 | 3.40 | 3.40 | 3.39 | 3.31 | 3.48 | 3.49 | 3.53 | 3.35 |
| RMSE $\tau_\phi$ | | 8.70 | 8.26 | 5.01 | 4.98 | 2.37 | 4.10 | 4.27 | 4.64 | 2.01 |
| RMSE $\tau_{\psi,k}$ | | 7.58 | 2.43 | 4.72 | 4.62 | 4.72 | 5.41 | 5.03 | 5.92 | 4.93 |
| RMSE $\log(\lambda_j)$ | | 62.33 | 50.86 | 38.17 | 38.03 | 0.00 | 31.53 | 32.13 | 34.70 | 0.00 |
| RMSE $\beta_{o,\eta,j}$ | | 4.07 | 4.64 | 4.45 | 4.52 | 5.16 | 4.30 | 4.32 | 1.25 | 0.00 |
| RMSE $\beta_{o,z,j}$ | $\times 100$ | 4.37 | 4.93 | 4.51 | 4.51 | 4.97 | 4.65 | 4.71 | 1.83 | 0.00 |
| RMSE $\beta_{\mathrm{D},\eta,k}$ | | 5.36 | 6.31 | 3.88 | 4.07 | 3.60 | 3.89 | 3.92 | 3.30 | 2.55 |
| RMSE $\beta_{\mathrm{D},z,k}$ | | 5.67 | 9.41 | 5.23 | 5.35 | 3.32 | 4.65 | 4.87 | 3.75 | 2.47 |
| RMSE OS $\eta$ | | 4.07 | 6.11 | 3.96 | 3.97 | 5.00 | 4.03 | 4.06 | 3.47 | 2.53 |
| RMSE OS $z$ | | 4.26 | 7.75 | 4.09 | 4.09 | 5.00 | 4.01 | 4.03 | 3.64 | 2.31 |

Table 2. Values of bias$\times$100 and RMSE$\times$100, for some particular treatments in the simulation study. In all of these cases, we use 4, 6, and 3 outcomes in the three domains, a large sample size ($n = 500$), the large exposure effect ($\bar{\beta}_{D,\eta} = 0.15$), and large deviations between domains (sd($\beta_{D,\eta}$) and sd($\beta_{D,z}$)). Also, recall the true values of $\tau_o = \tau_{o,1} = 0.05$, $\sigma_j = 1$, $\tau_\phi = 0.2$, and $\tau_{\psi,k} = 0.05$.

| Variable | Counts or Mean ±SD | Avg. Phthalate Score (By Category or $\leq$ / $>$ Variable Median) | Regression Coefficient for Phthalate Score ($\times 100$) |
|---|---|---|---|
| **Mother's Race** (Cauc. / Non-) | 89 / 29 | 7.53 / 7.07 | – |
| **Mother's Age** | $30.1 \pm 5.08$ | 8.46 / 6.37 | – |
| **Infant Age** (mos.) | $10.3 \pm 7.30$ | 8.02 / 6.81 | – |
| **Gestational Age** (wks.) | $39.0 \pm 2.18$ | 7.55 / 7.19 | – |
| **Creatinine** | $88.5 \pm 62.1$ | 4.29 / 10.54 | – |
| **Skinfold Thickness Flank** | $5.56 \pm 1.86$ | 7.26 / 7.60 | $-13.5\ (-40.3, 13.3)$ |
| **Skinfold Thickness Quadriceps** | $14.82 \pm 5.42$ | 6.89 / 8.19 | $5.38\ (-22.3, 33.1)$ |
| **Skinfold Thickness Subscapular** | $6.96 \pm 1.98$ | 7.04 / 8.12 | $-3.29\ (-28.2, 21.6)$ |
| **Skinfold Thickness Triceps** | $9.74 \pm 2.42$ | 7.00 / 7.86 | $13.1\ (-13.5, 39.8)$ |
| **Body Mass Index** | $16.8 \pm 1.49$ | 6.90 / 7.95 | $7.76\ (-20.7, 36.2)$ |
| **Weight Percentile** | $49.1 \pm 31.5$ | 6.90 / 7.93 | $11.1\ (-15.6, 37.7)$ |
| **Head Circumference Percentile** | $56.3 \pm 30.1$ | 7.53 / 7.31 | $-1.26\ (-29.4, 26.9)$ |

Table 3. Column 2: Summaries of the anthropometry measurements and covariates; Column 3: average phthalate score for individuals in each category (for categorical variables) or individuals above and below the variable median (for continuous variables); Column 4: regression coefficient for phthalate score, with 95% confidence interval.

| | Model A | Model C | Model D | Model E |
|---|---|---|---|---|
| ST Flank | $-3.32\ (-27.6, 19.8)$ | $-3.13\ (-25.6, 18.9)$ | $0.345\ (-19.0, 19.9)$ | $0.221\ (-13.5, 13.6)$ |
| ST Quadriceps | $-0.317\ (-20.3, 20.7)$ | $-2.01\ (-17.5, 13.6)$ | $0.345\ (-19.0, 19.9)$ | $0.221\ (-13.5, 13.6)$ |
| ST Subscapular | $-2.40\ (-24.5, 19.5)$ | $-3.27\ (-27.0, 20.2)$ | $0.345\ (-19.0, 19.9)$ | $0.221\ (-13.5, 13.6)$ |
| ST Triceps | $1.03\ (-16.3, 20.0)$ | $-1.56\ (-14.2, 11.4)$ | $0.345\ (-19.0, 19.9)$ | $0.221\ (-13.5, 13.6)$ |
| BMI | $9.47\ (-15.0, 34.0)$ | $7.95\ (-10.8, 27.4)$ | $9.86\ (-14.5, 33.4)$ | $9.43\ (-10.0, 29.1)$ |
| Weight | $9.52\ (-15.2, 34.8)$ | $11.1\ (-14.9, 37.5)$ | $9.86\ (-14.5, 33.4)$ | $9.43\ (-10.0, 29.1)$ |
| Head Circum. | $-1.40\ (-31.1, 28.4)$ | $-1.27\ (-30.7, 28.5)$ | $-0.920\ (-31.3, 29.2)$ | $-1.34\ (-30.5, 26.9)$ |

* ST: Skinfold Thickness

Table 4. Estimated phthalate exposure effect $(\beta_{o,\eta,j} + \lambda_j \beta_{\mathrm{D},\eta,d(j)})$ times 100 for each of the outcome variables $j$. 95% credible intervals are shown in parentheses.

| Covariate | Domain 1 | Domain 2 | Domain 3 |
|---|---|---|---|
| Sqrt. Creatinine | $1.32\ (-12.2, 14.7)$ | $0.405\ (-18.5, 19.5)$ | $5.88\ (-22.4, 34.5)$ |
| Infant Age | $\mathbf{-27.9}\ (\mathbf{-37.1}, \mathbf{-18.9})$ | $\mathbf{-19.2}\ (\mathbf{-32.0}, \mathbf{-6.36})$ | $16.1\ (-2.68, 34.9)$ |
| Mother's Age | $\mathbf{-11.6}\ (\mathbf{-21.0}, \mathbf{-2.51})$ | $0.133\ (-13.2, 13.4)$ | $9.97\ (-9.40, 29.5)$ |
| Gestational Age | $8.29\ (-0.601, 17.3)$ | $\mathbf{17.5}\ (\mathbf{4.62}, \mathbf{30.4})$ | $10.4\ (-8.27, 29.1)$ |
| Race (0:Cauc., 1:Non-) | $4.43\ (-4.62, 13.6)$ | $4.56\ (-8.10, 17.3)$ | $-12.3\ (-31.1, 6.42)$ |

Table 5. Estimated covariate effects $(\boldsymbol{\beta}_{o,z,j} + \lambda_j \boldsymbol{\beta}_{\mathrm{D},z,d(j)})$ times 100 in Model E for the phthalates data. 95% credible intervals are shown in parentheses.

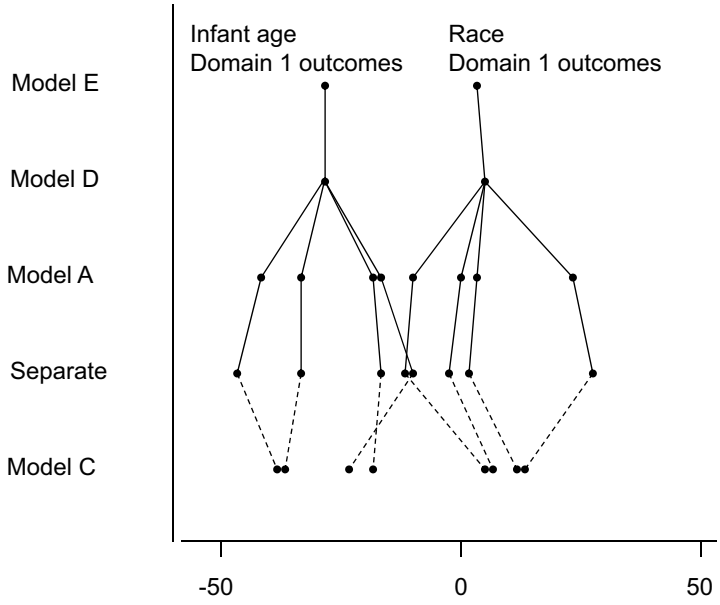


*Figure 1. Coefficient estimates ($\times 100$) for Models A, C, D, and E and separate regressions, for two representative predictors and the Domain 1 outcomes.*